Методы автоматического исправления ошибок в электронных документах

Манина Д.Р[1], Огнева М.В[2]
[1]manina.dasha@bk.ru, [2]ognevamv@gmail.com,
Саратовский государственный университет имени Н.Г.Чернышевского

На сегодняшний день особое внимание уделяется вопросу дистанционного обучения. В связи с этим, школьникам приходится больше работать с электронными документами и пользоваться поисковыми системами. Одна из проблем, которая при этом возникает − это проблема обнаружения и исправления ошибок при написании слов. В статье рассмотрены алгоритмы, которые используются для решения данной проблемы и приводится их анализ.

Ключевые слова: нечеткий поиск, расстояние Левенштейна, фонетический алгоритм, фонетическое расстояние.

Введение

В настоящее время особенно актуальным является вопрос о дистанционном обучении. В связи с этим, школьникам приходится больше работать с электронными документами и пользоваться поисковыми системами, для нахождения нужной информации при обучении.

Такие системы позволяют из множества текстов отбирать релевантные, соответствующие определённому запросу. Запрос представляет собой одно или несколько ключевых слов, которые, как предполагается, содержатся в искомом документе. На этом этапе начинают возникать сложности, так как в запросе могут содержаться разного рода ошибки, которые пользователь допустил по случайности, либо вследствие своей неграмотности. В результате поисковая система выдаст огромное количество ссылок, большинство из которых не отвечают запросу и являются информационным мусором.

Для предотвращения вышеперечисленных ситуаций используются алгоритмы фонетического кодирования. Такие алгоритмы устраняют для пользователя необходимость знать правильное написание каждого термина, с которым он работает и учитывают возможные ошибки и опечатки пользователей, допущенные ими при вводе.

Алгоритмы фонетического кодирования являются основой для построения современных систем проверки орфографии, которые используются в текстовых редакторах.

Также алгоритмы фонетического кодирования используются для функций, которые выдают пользователю сообщение «возможно вы имели в виду…» в поисковых системах, вроде Google или Yandex. При этом происходит поиск элементов, которые в наибольшей степени близки по написанию к запрошенному термину или фразе.

Алгоритмы фонетического кодирования

Алгоритмы фонетического кодирования разделены на алгоритмы для сравнения слов – фонетические алгоритмы и алгоритмы определения расстояния между словами – фонетические расстояния [1].

Фонетические алгоритмы представляют собой группировку слов со схожим произношением с помощью закодированной строки, в которую они преобразуются на основе последовательности букв слова и правил произношения. По закодированным строкам двух различных слов можно делать выводы о близости этих слов по звучанию, то есть смотреть насколько совпадают или близки получившиеся закодированные последовательности. Большинство фонетических алгоритмов предназначены для английского языка, но некоторые из них адаптированы и для других языков, в том числе и для русского. Для этого нужно, чтобы алгоритм учитывал правила фонетического кодирования языка и фонетические особенности языка.

Фонетическое расстояние применяется в алгоритмах нечеткого поиска. Оно определяет близость строк по написанию с помощью метрики – функции расстояния, которая сопоставляет двум строкам некоторое число, по которому можно судить об их различии. Происходит определение сходства слов по произношению путем подсчета расстояния между словами по написанию. Такие алгоритмы не берут во внимание языки запроса и множества элементов, по которым ведётся поиск – для них важна только операция сравнения символов. На данный момент не существует идеального варианта для исправления слов с помощью фонетических алгоритмов. Работы в данном направлении ведутся и по сей день, поэтому эта тема является актуальной.

Нами рассматривались аспекты отдельного и совместного использования фонетических алгоритмов Metaphone и Polyphone и фонетических расстояний Левенштейна и Дамерау-Левенштейна.

Для тестирования был взят словарь с русскими словами размером 100000 слов. Результаты работы представлены в таблице.

Таблица 1. – Результаты работы алгоритмов

Алгоритм	Точность исправления (%)	Среднее время выполнения (мс)
Расстояние Левенштейна	89,35	79536
Расстояние Дамерау–Левенштейна	90,95	83798
Metaphone	88,8	85762
Polyphone	94,35	202952
Metaphone & расстояние Дамерау–Левенштейна	96,55	151553
Polyphone & расстояние Левенштейна	95,5	217095

Таким образом, было выявлено, что самая высокая точность исправления ошибок у объединения фонетического алгоритма Metaphone и фонетического расстояния Дамерау-Левенштейна. Такое совмещение было выбрано неслучайно, оно позволяет использовать достоинства каждого алгоритма и уменьшить зависимость от недостатков. Например, в случае использования фонетических алгоритмов, близкие по звучанию слова могут иметь совершенно разный код. Это произойдет из-за того, что пользователь может пропустить букву или добавить лишнюю, что приведет к изменению закодированной строки. Также есть слова, которые, несмотря на свою фонетическую схожесть, не попадают в результирующий набор из-за слишком «строгих» правил алгоритмов. Поэтому без использования посимвольного сравнения элементов с помощью фонетических расстояний в конечном результате может не оказаться ни одного из релевантных элементов. Аналогично использование одних только фонетических расстояний не всегда эффективно, так как они не берут во внимание фонетику и аспекты языка. Кроме того, использованию фонетических расстояний присущи большие вычислительные затраты.

Описание программы

Для демонстрации работы гибридных алгоритмов, было разработано приложение на языке C# с использованием Windows Forms [2].

При запуске программы появляется окно, изображенное на рисунке 1. Для проверки правильности написания слова, пользователю предлагается ввести его в поисковую строку.

Для слова из поисковой строки и каждого слова в словаре вычисляется Metaphone-код. В качестве результата выводится слово или слова из словаря, для которых расстояние Дамерау–Левенштейна Metaphone-кода слова из поисковой строки и Metaphone-кода слова из словаря меньше 2. В противном случае считаем, что пользователь сделал слишком много ошибок в слове.

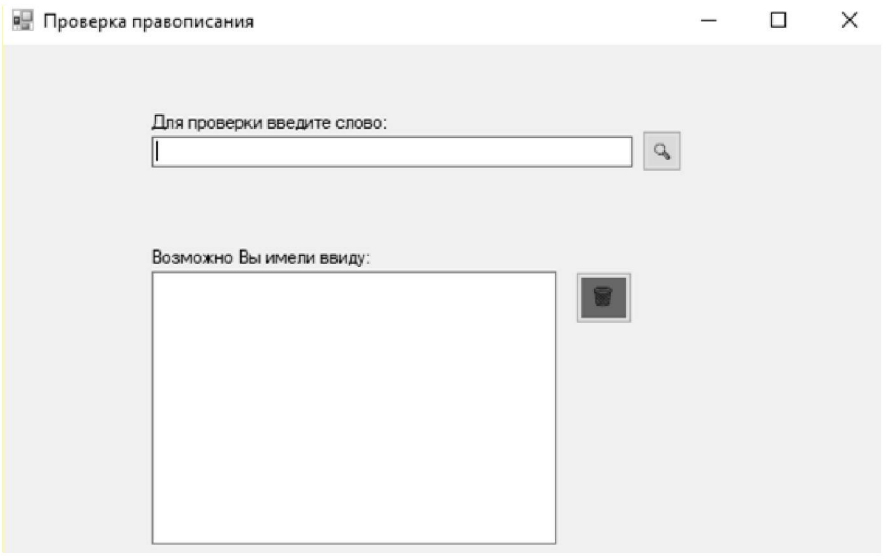

Рис. 1. Окно при запуске программы

В текстовом поле, изображенном на рисунке 2, выводятся все слова, на которые пользователю предлагается заменить ошибочное слово. Стоит заметить, что если слово из поисковой строки имеет небольшую длину, и пользователь допустил в нем ошибки, то при подсчете расстояния Дамерау-Левенштейна данному слову могут соответствовать несколько слов из словаря, расстояние до которых будет минимальным. Это означает, что имеются неоднозначные случаи исправления слова.

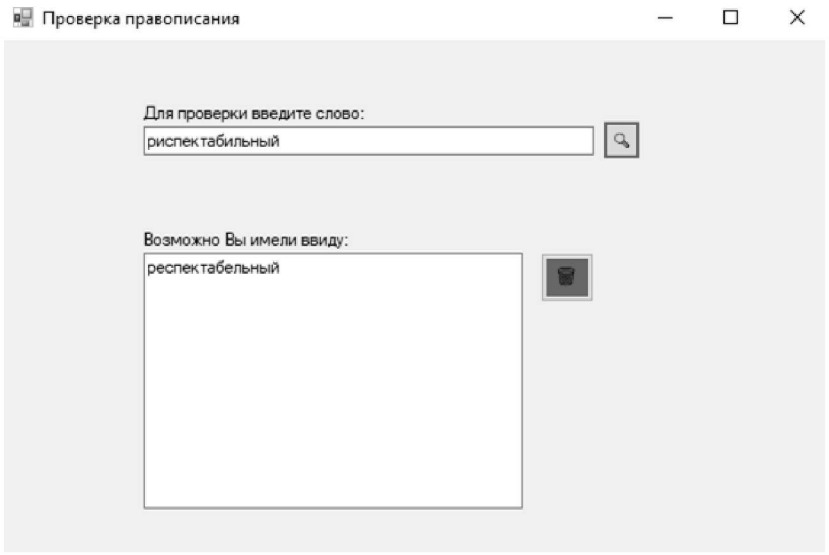

Рис. 2. Окно после проверки слова программой

Например, неправильно написанному слову «грам» соответствуют слова из словаря «грамм» и «гром», пользователь мог иметь ввиду любой из этих двух вариантов. Это означает, что имеется неоднозначный случай совпадения слов в диапазоне 2. На рисунке 3 представлен пример неоднозначного случая исправления слова.

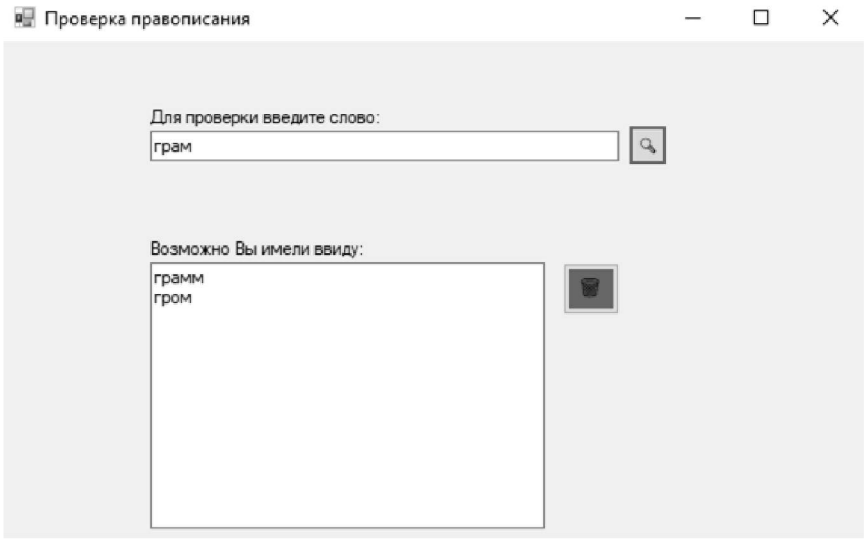

Рис. 3. Неоднозначный случай

Заключение

По результатам проведенных экспериментов можно сделать вывод, что совместная интеграция фонетических алгоритмов и фонетических расстояний, обеспечивает более эффективный результат исправления с точностью свыше 95%.

Это дает возможность проводить более глубокий и затрагиваемый различные аспекты анализ информации, что позволяет повысить качество научных исследований.

Список литературы

[1] *Выхованец В. С., Ду Ц., Сакулин С. А.* Обзор алгоритмов фонетического кодирования // Управление большими системами: сборник трудов, no. 73, 2018. С. 67–94.

[2] Windows Forms [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Windows_Form (дата обращения 18.09.2020).